Vítor Teófilo Rosemberg

Estudo sobre a Incidência de Acidentes Fatais no Trânsito do Distrito Federal

Brasil

2018, v-1.9.6

Vítor Teófilo Rosemberg

Estudo sobre a Incidência de Acidentes Fatais no Trânsito do Distrito Federal

Universidade de Brasília – UnB

Departamento de Estatística

Trabalho de Conclusão de Curso

Orientador: Profº Leandro Tavares Correia

Brasil

2018, v-1.9.6

Vítor Teófilo Rosemberg
Estudo sobre a Incidência de Acidentes Fatais no Trânsito do Distrito Federal/ Vítor Teófilo Rosemberg. – Brasil, 2018, v-1.9.6-
58 p. : il. (algumas color.) ; 30 cm.

Orientador: Profº Leandro Tavares Correia

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Departamento de Estatística
Trabalho de Conclusão de Curso, 2018, v-1.9.6.
1. Acidente. 2. Trânsito. 3. Regressão Logística. 4. Distrito Federal.

Vítor Teófilo Rosemberg

Estudo sobre a Incidência de Acidentes Fatais no Trânsito do Distrito Federal

Trabalho aprovado. Brasil, 28 de junho de 2018:

Profº Leandro Tavares Correia
Orientador

Banca
Profª Maria Teresa Leão Costa

Banca
Profº José Augusto Fiorucci

Brasil
2018, v-1.9.6

Este trabalho é dedicado às pessoas que se importam com o bem estar no trânsito e se preocupam com a vida.

Agradecimentos

Os agradecimentos principais são direcionados à Karina Teófilo Rosemberg, Roberto Rodriguez Rosemberg, Profº Leandro Tavares Correia e todos aqueles que contribuíram para que a produção desse trabalho acadêmico fosse possível.

Agradecimentos especiais são direcionados à Gerência de Estatística de Acidentes de Trânsito do DETRAN-DF[1] e ao Profº David Duarte Lima, especialista em educação para o trânsito [2].

[1] <http://www.detran.df.gov.br/>
[2] <http://ist.org.br/>

Resumo

Uma das maiores preocupações do Detran-DF é a morte no trânsito. Apesar do investimento na área de educação no trânsito, onde até o dia 31/08 deste ano foram investidos em torno de R$ 3,2 milhões, o número de mortes no trânsito continua elevado. Em uma estimativa conservadora, os acidentes de trânsito em rodovias custam à sociedade brasileira aproximadamente R$ 40 bilhões por ano. O principal objetivo deste estudo é mapear os acidentes de trânsito com morte no Distrito Federal e ver onde os mesmos estão mais propensos a ocorrer. Uma alternativa para atingir este objetivo é por meio da regressão logística, que é utilizada com o intuito de relacionar a variável resposta com as variáveis explicativas, sendo a variável resposta uma variável binária. Neste caso, a variável resposta seria dicotômica, sendo a ocorrência de fatalidade um "evento", por isso, a regressão logística seria uma boa opção para fazer este estudo. Para a situação proposta, talvez a regressão logística usual não seria suficiente para modelar os dados de maneira eficiente e muito menos para uma previsão correta.Existem alguns estudos que apontam que o modelo de regressão logística usual não funciona bem em situações que as caudas da distribuição das probabilidades estimadas são mais pesadas, como constatado por Stukel (1988) e também que o modelo logito usual subestima a probabilidade do evento de interesse quando o mesmo é construído utilizando bases com dados extremamente desbalanceados, apresentado por King e Zeng (2001).

Palavras-chave: acidente. trânsito. regressão logística. distrito federal.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Uma das maiores preocupações do Detran-DF é a morte no trânsito. Apesar do investimento na área de educação no trânsito, onde até o dia 31/08 deste ano foram investidos em torno de R$ 3,2 milhões, o número de mortes no trânsito continua elevado. Porém, este investimento não representa uma quantia expressiva, levando em conta do faturamento que o órgão fatura com as multas. (DETRAN, 2017)

A segurança no trânsito é de preocupação geral dos governos. Além de diminuir a perda de capital humano diminuindo os mortos e feridos no trânsito, diminui também os custos hospitalares públicos. A situação da segurança de trânsito no Brasil parece estar entre as mais graves no mundo, com cerca de 45 mil óbitos e centenas de milhares de feridos a cada ano (2012). No Distrito Federal não é diferente. (DETRAN, 2017)

Além deste problema da segurança no trânsito, existe a preocupação financeira que os acidentes de trânsito trazem para a sociedade. Em uma estimativa conservadora, os acidentes de trânsito em rodovias custam à sociedade brasileira aproximadamente R$ 40 bilhões por ano. Em áreas urbanas, a estimativa do custo é próximo de R$ 10 bilhões. Esse custo é consideravelmente mais elevado quando há fatalidade no acidente, aumentando substancialmente o custo final gerado. Essas estimativas levam em consideração gastos hospitalares, perda de produção, remoção do acidente, entre outras coisas. (IPEA, 2016)

A preocupação com os custos gerados pelos acidentes para a sociedade é um fator muito importante a ser levado em consideração, porém, a preocupação com a vida também deve ter uma atenção maior. Não se pode calcular o que representa a perda de uma vida humana ou os danos psíquicos e estresses traumáticos aos quais as vítimas de trânsito e seus familiares são submetidos após eventos dessa natureza. (LIMA; RODRIGUES, 2016)

O principal objetivo deste estudo é mapear os acidentes de trânsito com morte no Distrito Federal e ver onde os mesmos estão mais propensos a ocorrer. Uma alternativa para atingir este objetivo é por meio da regressão logística, que é utilizada com o intuito de relacionar a variável resposta com as variáveis explicativas, sendo a variável resposta uma variável binária. Neste caso, a variável resposta seria dicotômica, sendo a ocorrência de fatalidade um "evento", por isso, a regressão logística seria uma boa opção para fazer este estudo. As variáveis explicativas, ou seja, o que poderia

influenciar a presença ou não de fatalidade em um acidente de trânsito, seriam a idade e sexo do motorista, local do acidente, tipo de veículo, horário do dia, entre outros fatores.

Para a situação proposta a regressão logística usual não seria suficiente para modelar os dados de maneira eficiente e muito menos para uma previsão correta. Isso ocorre pelo fato da distribuição da variável resposta ser extremamente desbalanceada, ou seja, a proporção de acidentes fatais no Distrito Federal é muito baixa, não chegando a 4% do total dos acidentes. Existem alguns estudos que apontam que o modelo de regressão logística usual não funciona bem em situações que as caudas da distribuição das probabilidades estimadas são mais pesadas, como constatado por Stukel (1988) e também que o modelo logito usual subestima a probabilidade do evento de interesse quando o mesmo é construído utilizando bases com dados extremamente desbalanceados, segundo King e Zeng (2001).

Existem alternativas para contornar os problemas citados anteriormente. Uma delas é a técnica de amostragem que utiliza amostras *state-dependent*. Essa técnica, juntamente com o Modelo de Correção a Priori, que permite manter as propriedades dos estimadores de máxima verossimilhança quando utilizada esta amostragem, é muito utilizada para auxiliar no ajuste do modelo logito usual. (KING; ZENG, 2001)

Além da alternativa citada, existem outras opções para contornar este problema como o modelo logito limitado, sugerido por Cramer (2004), que estabelece um limite superior para a probabilidade de sucesso. Além das opções apresentadas para resolver o problema gerado pelo elevado desbalanceamento dos dados, existe a opção da utilização de um modelo logito generalizado, na qual muitos autores apresentaram propostas que generalizem o modelo logito padrão e a de um modelo logito com resto de origem, sugerido por Suissa, como mostrado por Scacabarozi (2012).

Para atingir o objetivo principal deste trabalho, será feita uma análise e interpretação de modelos logísticos, propor diferentes abordagens de modelos logísticos para baixas proporções de sucesso (eventos raros), em seguida fazer um estudo de simulação para verificação do melhor modelo logístico a ser utilizado (técnica mais eficiente) e por final, aplicação de dados reais (Dados do Detran-DF entre 2015 e 2016).

O estudo de simulação feito neste trabalho e as análises dos resultados foram feitos utilizando o software R 3.5.1.

2 Metodologia

Como falado anteriormente, a regressão logística usual não seria ideal para resolver os problemas do desbalanceamento dos dados que serão estudados, desta maneira, será feita uma revisão bibliográfica na parte de modelagem de eventos raros, mais especificamente, na melhor técnica para se analisar dados extremamente desbalanceados.

A técnica de amostragem *state-dependent* junto com o Método de Correção a Priori é uma alternativa, porém existem outras como o estimador para o vício de qualquer modelo linear generalizado proposto por McCullagh e Nelder (1989) que posteriormente foi adaptado por King e Zeng (2001) para o uso junto com a *state-dependent*. Outra alternativa é o modelo logito limitado, sugerido por Cramer (2004), que estabelece um limite superior para a probabilidade de sucesso.

Após essa revisão bibliográfica, será feito um estudo de simulação para definir a melhor técnica para solucionar este problema do desbalanceamento dos dados. Definida a melhor técnica, esta será aplicada nos dados reais fornecidos pela GEREST, Gerência de Estatística do Detran-DF, para os anos de 2015 e 2016, com o intuito de descobrir os principais fatores que influenciam a fatalidade em um acidente de trânsito e prever os locais mais propensos a se ter acidentes com vítimas fatais. O banco de dados fornecido pelo Detran-DF é construído a partir de uma operação conjunta do órgão com a Polícia Civil do DF[1]. Este banco possui diversas variáveis, entre elas, a localização do acidente, presença de vítima fatal, idade do condutor, dia e horário do acidente, tipo de veículo, entre várias outras.

Para maiores detalhes, consultar King e Zeng (2001), Stukel (1988) e Cramer (2004). Consultar também o site do Detran-DF[2] para melhores explicações sobre a coleta dos dados de acidentes.

[1] <http://www.detran.df.gov.br/wp-content/uploads/2018/06/SAT-Sistema_Informacoes_Acidente.pdf>

[2] <http://www.detran.df.gov.br/>

2.1 Regressão Logística Usual

Os métodos de regressão tem como objetivo descrever as relações entre a variável resposta $(\boldsymbol{Y})$ e a variável explicativa $(\boldsymbol{X})$. Em problemas de regressão, a quantidade de interesse é a esperança condicional de Y dado X, $\boldsymbol{E(Y|X=x)}$. Neste caso, essa esperança pode ser expressa como uma equação linear de $\boldsymbol{x}$:

$$\boldsymbol{E(Y|x) = \beta_0 + \beta_1 x} \tag{2.1}$$

A regressão logística é utilizada em situações que a variável resposta é dicotômica, ou binária, ou seja, pode assumir o valor 1 (*"sucesso"*) ou 0 (*"fracasso"*). Quando a variável resposta segue uma distribuição Bernoulli, como é o caso na regressão logística, sua média condicional $\boldsymbol{E(Y|X=x)}$ deve estar entre 0 e 1. De maneira geral, é utilizado $\boldsymbol{\pi_i = E(Y|X=x_i)}$ para representar a média condicional de $\boldsymbol{Y}$ dado $\boldsymbol{x_i}$ quando a distribuição logística é utilizada. (HOSMER; LEMESHOW, 2000)

É importante ressaltar que a regressão linear não é recomendada para esse caso particular (variável resposta dicotômica) pois, a suposição que os erros tem distribuição normal não é verificada neste caso.

2.1.1 Modelo

No modelo de regressão logística múltipla, têm-se que $\boldsymbol{Y_i|X_i}$ tem distribuição Bernoulli com parâmetro de sucesso $\boldsymbol{\pi_i}$, onde $\boldsymbol{Y_i}$ é a $\boldsymbol{i}$-ésima variável resposta e $\boldsymbol{X_i}$ o vetor de covariáveis do $\boldsymbol{i}$-ésimo indivíduo. Para resumir, seja $\boldsymbol{x_i}$ uma observação de $\boldsymbol{X_i}$, então a distribuição $\boldsymbol{Y_i|X_i}$ é dada por:

$$\boldsymbol{Y_i|X_i = x_i \sim Ber(\pi(x_i))} \tag{2.2}$$

A partir de agora será utilizada a notação $\boldsymbol{\pi_i}$ ao invés de $\boldsymbol{\pi(x_i)}$ a fim de facilitar a escrita. A função distribuição de probabilidade de $\boldsymbol{Y_i|X_i = x_i}$ é:

$$\boldsymbol{P(Y_i|x_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i},} \tag{2.3}$$

com $\boldsymbol{y_i = 0, 1}$ e $\boldsymbol{i = 1, \ldots, n}$

A observação $\boldsymbol{x_i}$ representa o vetor coluna de $\boldsymbol{p}$ covariáveis observadas do i-ésimo indivíduo. Normalmente, o primeiro elemento de $\boldsymbol{x_i}$ vale 1, com a finalidade de permitir que o modelo tenha intercepto $\boldsymbol{(\beta_0)}$ não nulo. Segundo Hosmer e Lemeshow

(2000), o **logito do modelo de regressão logística múltipla** é dado pela seguinte equação:

$$g(\pi_i) = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \cdots + \beta_p x_p = x_i'\beta, \tag{2.4}$$

Onde, $x_i = (x_1, \ldots, x_p)$ e $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$. Dessa maneira, o **modelo da regressão logística múltipla** é dado por:

$$\pi_i = \frac{e^{g(\pi_i)}}{1 + e^{g(\pi_i)}}, i = 1, \ldots, n \tag{2.5}$$

onde π_i é a **probabilidade de sucesso** para o i-ésimo indivíduo. Ou seja, $P(Y = 1|x_i) = \pi_i$ e $g(.)$ é chamada função de ligação.

2.1.2 Estimação

Como falado na seção subseção 2.1.1, a distribuição de probabiliadde de $Y_i|X_i = x_i$ pode ser escrita como:

$$P(Y_i|x_i) = \pi^{y_i}(1 - \pi_i)^{1-y_i},$$

com $y_i = 0, 1$ e $i = 1, ..., n$

Assumindo independência entre as observações, podemos definir a função de máxima verossimilhança do modelo. Essa função é dada por:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \tag{2.6}$$

Aplicando o logaritmo na função de máximo verossimilhança e substituindo os valores de π_i temos:

$$l(\beta) = \log(L(\beta)) = \sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)] \tag{2.7}$$

O valor de β que maximiza Equação 2.7 é o estimador de máxima verossimilhança. Para acharmos este β é necessário derivar Equação 2.7 com relação a $\beta = (\beta_0, \ldots, \beta_p)^T$ e igualar a zero, porém as equações resultantes deste processo são não-lineares e este estimador é obtido por meio de métodos numéricos de otimização. Denotaremos $\hat{\beta}$ como o estimador de máxima verossmilhança de β.

Sabemos que $\hat{\beta}$, por ser um estimador de máxima verossmilhança, é consistente e assintoticamente eficiente e possui matriz de variâncias e covariâncias dada por:

$$V(\hat{\beta}) = \left[\sum_{i=1}^{n} \pi_i(1 - \pi_i)x_i x_i'\right]^{-1} \tag{2.8}$$

2.1.2.0.1 Vício

O vício do estimador do vetor de parâmetros de qualquer MLG (Modelo Linear Generalizado) é dado por: (MCCULLAGH; NELDER, 1989)

$$\boldsymbol{vicio}(\hat{\beta}) = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'W\xi} \tag{2.9}$$

onde,

- $\boldsymbol{X'WX}$ é a matriz de informação de Fisher;
- $\boldsymbol{\xi_i} = \boldsymbol{-0.5\mu_i''/\mu_i'Q_{ii}}$
  - $\boldsymbol{\mu_i}$ é a inversa da função de ligação;
  - $\boldsymbol{\mu_i'}$ e $\boldsymbol{\mu_i''}$ são as derivadas de primeira e segunda ordem de $\boldsymbol{\mu_i}$;
  - $\boldsymbol{Q_{ii}}$ é o i-ésimo elemento da diagonal principal de $\boldsymbol{X(X'W'X)X'}$;

  Temos que:

$$\frac{\boldsymbol{\mu_i''}}{\boldsymbol{\mu_i'}} = \left(\frac{\mathbf{1} - \mathbf{exp}(\boldsymbol{\eta_i})}{\mathbf{1} + \mathbf{exp}(\boldsymbol{\eta_i})}\right)$$

Onde, $\boldsymbol{\eta_i}$ é o preditor linear: $\boldsymbol{\eta_i} = \boldsymbol{x_i'\beta}$.

2.1.3 Interpretação dos Parâmetros

A interpretação dos parâmetros na regressão que tem a função de ligação a logito, a regressão logística, fornece uma interpretação conveniente dos parâmetros. Saber o impacto que as covariáveis causam na determinação da probabilidade de sucesso do evento de interesse é fundamental.

O modelo de Bernoulli permite o ajuste de algumas funções de ligação, como a Probit, log-log complementar, logito, entre outras. A ligação logito é a mais comumente utilizada e fornece uma interpretação conveniente dos parâmetros.

Cada função de ligação possui a sua interpretação natural dos coeficientes de regressão. Na regressão probit, a interpretação natural é o Risco Relativo, na log-log complementar, a Dose Letal. Já na logito, a interpretação natural dos coeficientes é dada pela *Odds Ratio*, também conhecida como Razão de Chances.

A chance de ocorrer um evento ($\boldsymbol{x_1}$) é dada por:

$$\frac{\boldsymbol{\pi_i}}{\mathbf{1} - \boldsymbol{\pi_i}}$$

Onde, aplicando na fórmula em Equação 2.4, temos:

$$\frac{\boldsymbol{\pi_i}}{\mathbf{1}-\boldsymbol{\pi_i}} = \exp \boldsymbol{g}(\boldsymbol{\pi_i}) = \exp(\boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{i1}} + \cdots + \boldsymbol{\beta_p x_{ip}})$$

Se adicionarmos, em uma unidade, uma variável independente contínua $\boldsymbol{x_1}$, mantendo todas as demais constantes, a chance do evento$(\boldsymbol{x_1}+\mathbf{1})$ é:

$$\begin{aligned}\frac{\boldsymbol{\pi_i^*}}{\mathbf{1}-\boldsymbol{\pi_i^*}} &= \exp(\boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x_{i1}}+\mathbf{1}) + \cdots + \boldsymbol{\beta_p x_{ip}}) \\ &= \exp(\boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{i1}} + \cdots + \boldsymbol{\beta_p x_{ip}}) \exp(\boldsymbol{\beta_1}) \\ &= \exp \boldsymbol{g}(\boldsymbol{\pi_i}) \exp(\boldsymbol{\beta_1})\end{aligned}$$

Desta maneira, a *Odds Ratio*$(\boldsymbol{\Psi})$ de $\boldsymbol{x_1}+\mathbf{1}$ em relação a $\boldsymbol{x_1}$ é dada por:

$$\begin{aligned}\boldsymbol{\Psi}(\boldsymbol{x_1}+\mathbf{1}, \boldsymbol{x_1}) &= \frac{\boldsymbol{\pi_i^*}/(\mathbf{1}-\boldsymbol{\pi_i^*})}{\boldsymbol{\pi_i}/(\mathbf{1}-\boldsymbol{\pi_i})} \\ &= \frac{\exp \boldsymbol{g}(\boldsymbol{\pi_i}) \exp(\boldsymbol{\beta_1})}{\exp \boldsymbol{g}(\boldsymbol{\pi_i})} \\ &= \exp(\boldsymbol{\beta_1})\end{aligned}$$

Interpretando este resultado, a chance do evento ocorrer entre os indivíduos que na variável $\boldsymbol{x_1}$ diferem em uma unidade é $\boldsymbol{e^{\beta_1}}$.

A estimativa da *Odds Ratio*, ou Razão de Chances, é dada por:

$$\widehat{\boldsymbol{\Psi}}(\boldsymbol{x_1}+\mathbf{1}, \boldsymbol{x_1}) = \exp(\widehat{\boldsymbol{\beta_1}})$$

De forma geral, não é necessário que esse acréscimo seja de apenas uma unidade, podendo a *Odds Ratio* ser estimada com acréscimo de c unidades. Sendo assim, substituindo $(\boldsymbol{x_1}+\mathbf{1})$ por $(\boldsymbol{x_1}+\boldsymbol{c})$, temos que a estimativa da Razão de Chances é:

$$\widehat{\boldsymbol{\Psi}}(\boldsymbol{x_1}+\boldsymbol{c}, \boldsymbol{x_1}) = \exp(\boldsymbol{c}\widehat{\boldsymbol{\beta_1}})$$

Na tentativa de facilitar a compreensão é proposto o seguinte exemplo: Seja $\boldsymbol{y}$ a variável que denota a presença de fatalidade no acidente de carro e $\boldsymbol{x}$ uma variável indicadora que denota se o motorista é do sexo masculino $(\boldsymbol{x}=\mathbf{1})$ ou do sexo feminino $(\boldsymbol{x}=\mathbf{0})$. Para simplificar, vamos assumir que existe apenas essa variável indicadora independente. E considerando as probabilidades de sucesso $\boldsymbol{\pi(0)}$ e $\boldsymbol{\pi(1)}$

respectivas ao motorista ser do sexo feminino ou masculino, respectivamente, temos que a razão de chances é dada por:

$$\boldsymbol{\Psi(1,0) = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = \exp \beta_1}$$

Se $\boldsymbol{\Psi(1,0) = 2}$, conclui-se que a chance de ter uma fatalidade em um acidente de carro com o motorista do sexo masculino é duas vezes maior do que com motorista do sexo feminino.

2.1.4 Predição

Neste capítulo serão descritas medidas de qualidade da discriminação do modelo que tem como objetivo avaliar o poder preditivo do mesmo. Essas medidas são: sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e acurácia. Essas medidas são conhecidas como medidas de desempenho. Também é utilizada a estatística Kolmogorov-Smirnov para testar se as populações do evento e não-evento estão bem discriminadas em relação às probabilidades de sucesso.

2.1.4.1 Curva ROC

A Curva ROC (*Receiver Operating Characteristic*) é um método para avaliar graficamente a qualidade de modelos de classificação. Esse método é muito utilizado quando existe um desbalanceamento entre as classes.

Esta metodologia é frequentemente utilizada para determinação de um ponto de corte. Este ponto de corte (c), que é um número entre 0 e 1, é utilizado para limitar a classificação de uma observação como sucesso ou fracasso (evento ou não-evento). Definido o ponto de corte c pela curva ROC, cada indivíduo é classificado pela sua probabilidade de sucesso estimada quando comparada com o ponto de corte. Ou seja, se a probabilidade estimada do indivíduo for maior que o ponto de corte, este indivíduo é classificado como sucesso, caso contrário, como fracasso. Vale lembrar que a escolha do melhor ponto de corte (c) depende do estudo que está sendo realizado.

A Curva ROC é um gráfico composto pela sensibilidade (eixo y *vertical*) e pelo complementar da especificidade (eixo x horizontal).

Para cada c, calcula-se as medidas preditivas como a sensibilidade e especificidade. Essas medidas preditivas de desempenho serão explicadas a seguir:

2.1.4.2 Medidas Preditivas

Construído o modelo, é necessário avaliar se o mesmo possui capacidade de distinguir as observações em que há ou não ocorrência do evento de interesse. Obter as medidas de desempenho é um importante procedimento na avaliação do poder de predição do modelo.

Seja $\hat{\boldsymbol{Y}} = \mathbf{1}$ um indivíduo classificado como sucesso e $\hat{\boldsymbol{Y}} = \mathbf{0}$ caso o indivíduo seja classificado como fracasso. As definições para as medidas de desempenho são:

2.1.4.2.1 Sensibilidade (Sen)

A sensibilidade é a probabilidade do modelo classificar corretamente um sucesso (evento).

$$\mathbf{Sen} = \boldsymbol{P}(\hat{\boldsymbol{Y}} = \mathbf{1} | \boldsymbol{Y} = \mathbf{1})$$

2.1.4.2.2 Especificadade (Esp)

A especificidade é a probabilidade do modelo classificar corretamente um fracasso (não-evento).

$$\mathbf{Esp} = \boldsymbol{P}(\hat{\boldsymbol{Y}} = \mathbf{0} | \boldsymbol{Y} = \mathbf{0})$$

2.1.4.2.3 Valor Preditivo (VP)

Valor Preditivo Positivo (VPP): É a probabilidade de um indivíduo ser sucesso (evento), uma vez que o modelo o classificou como tal.

$$\mathbf{VPP} = \boldsymbol{P}(\boldsymbol{Y} = \mathbf{1} | \hat{\boldsymbol{Y}} = \mathbf{1})$$

Valor Preditivo Negativo (VPN): É a probabilidade de um indivíduo ser fracasso (não-evento), uma vez que o modelo o classificou como tal.

$$\mathbf{VPN} = \boldsymbol{P}(\boldsymbol{Y} = \mathbf{0} | \hat{\boldsymbol{Y}} = \mathbf{0})$$

2.1.4.2.4 Acurácia (ACC)

A acurácia é a probabilidade do modelo classificar corretamente um sucesso ou um fracasso.

$$\mathbf{ACC} = \boldsymbol{P}(\boldsymbol{Y} = \mathbf{1}, \hat{\boldsymbol{Y}} = \mathbf{1}) + \boldsymbol{P}(\boldsymbol{Y} = \mathbf{0}, \hat{\boldsymbol{Y}} = \mathbf{0})$$

2.1.4.2.5 Estimativas

Definidas as medidas de desempenho, é necessário estimar as probabilidades apresentar acima a partir do ponto de corte (c) escolhido para o estudo. Essas estimativas são dadas por:

- **VP - Verdadeiro Positivo**: observações que foram corretamente classificadas como sucesso (evento).
- **VN - Verdadeiro Negativo**: observações que foram corretamente classificadas como fracasso (não-evento).
- **FP - Falso Positivo**: observações que foram classificadas incorretamente como sucesso (evento).
- **FN - Falso Negativo**: observações que foram classificadas incorretamente como fracasso (não-evento).

Com isso, temos que as estimativas para as medidas de desempenho são:

Medidas	Definição	***Estimativa***
$\boldsymbol{Sen}$	$\boldsymbol{P(\hat{Y}=1\|Y=1)}$	$\frac{VP}{VP+FN}$
$\boldsymbol{Esp}$	$\boldsymbol{P(\hat{Y}=0\|Y=0)}$	$\frac{VN}{VN+FP}$
$\boldsymbol{VPP}$	$\boldsymbol{P(Y=1\|\hat{Y}=1)}$	$\frac{VP}{VP+FP}$
$\boldsymbol{VPN}$	$\boldsymbol{P(Y=0\|\hat{Y}=0)}$	$\frac{VN}{VN+FN}$
$\boldsymbol{ACC}$	$\boldsymbol{P(Y=1,\hat{Y}=1)+P(Y=0,\hat{Y}=0)}$	$\frac{VP+VN}{VP+VN+FB+FN}$

2.1.5 Seleção de Variáveis

A escolha do melhor modelo possível varia bastante de acordo com a situação estudada, mas em geral, quer-se escolher o modelo com maior parcimônia que explica bem os dados. (HOSMER; LEMESHOW, 2000)

Um dos métodos mais utilizados para seleção de variáveis em regressão logística é o método stepwise. Como este método é uma mistura de dois outros, *Backward* e *Forward*, estes dois métodos serão explicados abaixo:

2.1.5.0.1 Método *Forward*

O método *Forward* se baseia em um modelo inicial apenas com o intercepto. As variáveis são adicionadas, uma de cada vez e é testado se o coeficiente desta variável é diferente ou não de zero. Caso seja, passa para o próximo passo, adicionando outra variável e assim por diante, até ter passado por todas as variáveis.

2.1.5.0.2 Método *Backward*

O método *Backward* é o oposto do método *Forward*. O modelo inicial contém todas as variáveis explicativas. A retirada é baseada, assim como no método forward, na estatística F.

2.1.5.0.3 Método *Stepwise*

O método *Stepwise* é um método de seleção automática de variáveis explicativas para o modelo em estudo. Este método é uma combinação de dois outros métodos, o *Backward* e o *Forward*. Sem entrar em detalhe nestes dois métodos, o método *Stepwise* é um algoritmo que mistura a inclusão e a exclusão de variáveis de acordo com a sua importância, ou seja, inclui e exclui variáveis para ver qual modelo fica melhor. Existem diferentes critérios para considerar a importância da variável, uma delas é pelo nível de significância do teste da razão de verossimilhanças entre os modelos que incluem ou excluem as variáveis estudadas. (PAULA, 2013)

Como falado anteriormente, o método *Stepwise* é uma mistura entre os métodos *Backward* e *Forward*. Sendo assim, o modelo inicial é composto apenas pelo intercepto e duas variáveis são adicionadas e é testado se a primeira fica ou não no modelo. Isso ocorre até não existirem mais variáveis explicativas. (PAULA, 2013)

2.1.6 Qualidade do Ajuste

2.1.6.1 AIC e BIC

Nesta seção serão apresentados dois critérios de avaliação da qualidade do ajuste de modelos de regressão. Em geral, utiliza-se os valores destes critérios para

selecionar qual modelo melhor explica o fenômeno estudado. Esses critérios são o AIC (*Akaike's Information Criterion*) e o BIC (Critério de Informação Bayesiano). O AIC, proposto por Akaike em 1974, segundo Paula (2013), utiliza a Informação de Kullback-Leibler para verificar a adequabilidade do modelo estudado. Em geral, caso haja empate nos valores de um desses critérios, a desviância é utilizada para desempatar e funciona do mesmo jeito do AIC e BIC, quanto menor, melhor é o modelo.

Sendo assim, o cálculo do AIC é dado da seguinte maneira:

$$\boldsymbol{AIC = -2\log(L(\hat{\theta})) + 2p} \tag{2.10}$$

onde, $\boldsymbol{L(\hat{\theta})}$ é a verossimilhança do modelo e $\boldsymbol{p}$ é o número de parâmetros do mesmo.

Já o BIC, proposto por Schwartz em 1978, segundo Paula (2013) é calculado da seguinte maneira:

$$\boldsymbol{BIC = -2\log(L(\hat{\theta})) + p\log(n)} \tag{2.11}$$

Sendo $\boldsymbol{n}$ o tamanho da amostra.

2.1.6.2 Deviance

A qualidade do ajuste pode também ser verificada através do *deviance*. Lembrando do log da função de verossimilhança do modelo, na seção 3.1.2, temos que o *deviance* é uma relação entre o máximo desta função para o modelo de interesse e o máximo do log da função de verossimilhança para o modelo mais completo. Ou seja,

$$\boldsymbol{Deviance = -2[L_m - Ls]} \tag{2.12}$$

onde,

- $\boldsymbol{L_s}$: Logaritmo da função de verossimilhança do modelo saturado (n parâmetros).
- $\boldsymbol{L_m}$: Logaritmo do modelo em investigação (p parâmetros), onde p<n.

O *deviance* possui distribuição $\boldsymbol{\chi^2}$ com $\boldsymbol{n - k}$ graus de liberdade e a hipótese que é testada é que todos os parâmetros que estão no modelo saturado, mas não estão no modelo investigado são nulos.

2.1.6.3 Teste de Hosmer e Lemeshow

Segundo Hosmer e Lemeshow (2000), há uma estatística para analisar a qualidade do ajuste. Esta estatística se baseia na comparação dos números observados

e esperados pelo modelo em estudo. É sugerida a separação em g=10 grupos, divididos pelas probabilidades ajustadas, o primeiro com as menores e o último com as maiores. A estatística deste teste é dada por:

$$\hat{C} = \sum_{i=1}^{g} \frac{\left(O_i - n'_i \tilde{\pi}_i\right)^2}{n'_i \tilde{\pi}_i \left(1 - \tilde{\pi}_i\right)}, \tag{2.13}$$

onde

- $\tilde{\pi}_i = \frac{1}{n'_i} \sum_{j=n'_1+\cdots+n'_{i-1}+1}^{n'_1+\cdots+n'_i} \hat{\pi}(j)$, $\hat{\pi}(j)$ = j-ésima probabilidade ajustada;
- O_i, número observado de sucessos no i-ésimo grupo;
- n'_i, quantidade de elementos no i-ésimo grupo;

Através de estudos de simulação, observou-se que $\hat{C}$ apresenta distribuição assintótica nula bem próxima da distribuição qui-quadrado com (g-2) graus de liberdade. (HOSMER; LEMESHOW, 2000)

2.2 Adaptação para Regressão Logística Usual

Como falado anteriormente, a regressão logística não funciona muito bem em situações que as caudas da distribuição de probabilidades estimadas são mais pesadas, como proposto por Stukel (1988) e também que este modelo usual subestima a probabilidade do evento de interesse quando o mesmo é construído utilizando bases de dados extremamente desbalanceadas, segundo King e Zeng (2001). Sendo assim, nesta seção serão abordadas algumas alternativas para contornar essa situação.

2.2.1 Amostras *State-Dependent*

As amostras *state-dependent* se baseiam na seleção de uma amostra contendo todos os sucessos (eventos) contidos na base de dados original e selecionar, via amostra aleatória simples, um número de fracassos (não-eventos) igual ou superior ao número de sucessos, porém essa quantidade de fracassos deve ser inferior a quantidade de fracassos total da base de dados original.

As amostras geradas via *state-dependent* são muito utilizadas em diversas áreas, porém elas sozinhas não conseguem resolver o problema da subestimação da probabilidade do evento de interesse. Desta maneira, é necessário uma validação inferencial desta técnica para os parâmetros obtidos por meio desta técnica de amostragem e algumas adaptações também são necessárias.

2.2.1.1 Método de Correção a Priori

O método de correção a priori consiste em modelar normalmente a regressão logística usual aos dados, calculando os estimadores por máxima verossimilhança, porém ao mesmo tempo, corrigir essas estimativas com base em informações prévias, informações a priori.

Segundo este método, os $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ estimadores de máxima verossimilhança do modelo são consistentes e eficientes, porém para que o estimador de $\boldsymbol{\beta_0}$ também seja consistente, é necessário corrigi-lo:

$$\hat{\boldsymbol{\beta}}_{\mathbf{0}} - \boldsymbol{ln}\left[\left(\frac{\mathbf{1}-\boldsymbol{\tau}}{\boldsymbol{\tau}}\right)\left(\frac{\overline{\boldsymbol{y}}}{\mathbf{1}-\overline{\boldsymbol{y}}}\right)\right] \tag{2.14}$$

onde, $\boldsymbol{\tau}$ representa a proporção de sucessos na população e $\overline{\boldsymbol{y}}$ a proporção de sucessos na amostra. Nota-se, neste caso, que a população retrata a base de dados completa e a amostra retrata a base de dados após a amostragem state-dependent.

Por ser necessário corrigir apenas o intercepto, este método é simples e facilmente aplicado.

2.2.1.2 Estimadores KZ

O estimador para $\hat{\boldsymbol{\beta}}$ para $\boldsymbol{\beta}$ é viciado mesmo quando o tamanho da amostra é grande quando os dados são extremamente desbalanceados e mesmo que haja a correção deste estimador pelo seu vício estimado, a probabilidade de sucesso é viciada para $\boldsymbol{\pi_i}$. Sendo assim, é necessário corrigir os estimadores e as probabilidades estimadas.

Os estimadores KZ levam em consideração um processo de "pesagem"para o cálculo dos parâmetros e correção das probabilidades estimadas. Esse processo altera o cálculo do função de verossimilhança, maximizando agora, para achar os estimadores, a função log de verossimilhança com peso, sendo essa equação apresentada abaixo:

$$\begin{aligned}\ln \boldsymbol{L}\left(\boldsymbol{\beta}|\boldsymbol{y}\right) &= \boldsymbol{\omega_1}\sum_{\boldsymbol{Y_i=1}}\ln(\boldsymbol{\pi}_i) + \boldsymbol{\omega_0}\sum_{\boldsymbol{Y_i=0}}\ln(\mathbf{1}-\boldsymbol{\pi}_i)\\ &= -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{\omega_i ln}\left(\mathbf{1}+\boldsymbol{e}^{(\mathbf{1}-\mathbf{2}\boldsymbol{y_i})\boldsymbol{x_i\beta}}\right)\end{aligned}$$

Sendo $\boldsymbol{\omega_i} = \boldsymbol{\omega_1 Y_i} + \boldsymbol{\omega_0}\left(\mathbf{1}-\boldsymbol{Y_i}\right)$, onde $\boldsymbol{\omega_0} = (\mathbf{1}-\boldsymbol{\tau})/(\mathbf{1}-\overline{\boldsymbol{y}})$ e $\boldsymbol{\omega_1} = \boldsymbol{\tau}/\overline{\boldsymbol{y}}$.

2.2.1.2.1 Parâmetros

O cálculo do vício do estimador do vetor de parâmetros segue o trabalho de McGullagh e Nelder, citado em Equação 2.9, porém com adaptações para eventos

raros, adicionando o fator dos pesos no cálculo dos vícios. Sendo assim, o vício é definido por:

$$\boldsymbol{vicio}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'W\xi} \tag{2.15}$$

onde,

- $\boldsymbol{\xi_i} = \mathbf{0.5}\boldsymbol{Q_{ii}}\left[(\mathbf{1} + \boldsymbol{\omega_1}\hat{\boldsymbol{\pi}}_i - \boldsymbol{\omega_1})\right]$;
- $\boldsymbol{Q_{ii}}$ é a diagonal de $\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X'WX})^{-1}\boldsymbol{X'}$;
- $\boldsymbol{W} = \boldsymbol{diag}\left[\hat{\boldsymbol{\pi}}_i\left(\mathbf{1} - \hat{\boldsymbol{\pi}}_i\right)\boldsymbol{\omega}_i\right]$;

O estimador de $\boldsymbol{\beta}$ corrigido pelo vício é dado por:

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{vicio}(\hat{\boldsymbol{\beta}}) \tag{2.16}$$

A matriz de variâncias e covariâncias de $\tilde{\boldsymbol{\beta}}$ é aproximadamente $\left(\frac{n}{n+p-1}\right)^{\mathbf{2}}\boldsymbol{V}(\hat{\boldsymbol{\beta}})$. Temos que $\boldsymbol{V}(\tilde{\boldsymbol{\beta}}) < \boldsymbol{V}(\hat{\boldsymbol{\beta}})$, uma vez que $\left(\frac{n}{n+p-1}\right)^{\mathbf{2}} < \mathbf{1}$. Com isso, temos que a diminuição da variância dos estimadores é causada pela diminuição do vício deles. (MCCULLAGH; NELDER, 1989)

2.2.1.2.2 Probabilidades Estimadas

Se levarmos em consideração obtidos anteriormente sobre os estimadores de $\boldsymbol{\beta}$ temos que seria melhor usarmos $\tilde{\boldsymbol{\pi}}(\boldsymbol{x_i})$ ao invés de $\hat{\boldsymbol{\pi}}(\boldsymbol{x_i})$, pois $\tilde{\boldsymbol{\beta}}$ é menos viciado do que $\hat{\boldsymbol{\beta}}$ e $\boldsymbol{V}(\tilde{\boldsymbol{\beta}}) < \boldsymbol{V}(\hat{\boldsymbol{\beta}})$. Porém, este estimador não leva em consideração a incerteza em relação a $\boldsymbol{\beta}$. Isso pode acarretar em geração de estimativas viesadas da probabilidade de sucesso. (KING; ZENG, 2001)

Para contornar essa situação da incerteza em relação a estimação do modelo, temos que:

$$\boldsymbol{\pi}(\boldsymbol{x_i}) = \boldsymbol{P}(\boldsymbol{Y_i} = \mathbf{1}) = \int \boldsymbol{P}(\boldsymbol{Y_i} = \mathbf{1}|\boldsymbol{\beta^*})\boldsymbol{P}(\boldsymbol{\beta^*})\boldsymbol{d\beta},$$

onde, P(.) representa a incerteza com relação à $\boldsymbol{\beta}$. Usa-se a distribuição a posteriori de $\boldsymbol{\beta}$ como:

$$\boldsymbol{\beta} \sim \boldsymbol{Normal}\left[\beta|\tilde{\boldsymbol{\beta}}, \boldsymbol{V}(\tilde{\boldsymbol{\beta}})\right]$$

Podemos escrever $\boldsymbol{\pi}(\boldsymbol{x_i})$ expandindo em série de Taylor a expressão $\boldsymbol{\pi}(\boldsymbol{x_0})$ em torno de $\tilde{\boldsymbol{\beta}}$ até a segunda ordem, obtendo a seguinte expressão:

$$\boldsymbol{P}(\boldsymbol{Y_0} = \mathbf{1}) \approx \tilde{\boldsymbol{\pi}}(\boldsymbol{x_0}) + \left[\frac{\boldsymbol{\partial\pi}(\boldsymbol{x_0})}{\boldsymbol{\partial\beta}}\right]_{\beta=\tilde{\beta}}(\beta-\tilde{\beta}) + \frac{\mathbf{1}}{\mathbf{2}}(\beta-\tilde{\beta})\left[\frac{\boldsymbol{\partial^2\pi}(\boldsymbol{x_0})}{\boldsymbol{\partial\beta'\partial\beta}}\right]_{\beta=\tilde{\beta}}(\beta-\tilde{\beta}),$$

onde,

- $\boldsymbol{\pi(x_0) = \frac{e^{x_0'\beta}}{1+e^{x_0'\beta}}}$,
- $\boldsymbol{\left[\frac{\partial\pi(x_0)}{\partial\beta}\right]_{\beta=\tilde{\beta}} = \tilde{\pi}(x_0)(1-\tilde{\pi}(x_0))x_0'(\beta-\tilde{\beta})}$,
- $\boldsymbol{\left[\frac{\partial\pi(x_0)}{\partial\beta'\partial\beta}\right]_{\beta=\tilde{\beta}} = (0,5-\tilde{\pi}(x_0))\tilde{\pi}(x_0)(1-\tilde{\pi}(x_0))x_0'Dx_0}$
  - D é uma matriz de ordem k x k e o seu elemento k,j é dado por $(\beta_k-\tilde{\beta}_k)(\beta_j-\tilde{\beta}_j)$

Podemos escrever a forma aproximada de $\boldsymbol{\pi(x_i)}$ como:

$$\boldsymbol{P(Y_0=1) = E\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right) \approx}$$

$$\boldsymbol{\tilde{\pi}(x_0)+\tilde{\pi}(x_0)(1-\tilde{\pi}(x_0))x_0'b+(0,5+\tilde{\pi}(x_0))(\tilde{\pi}(x_0)-\tilde{\pi}^2(x_0))x_0'[V(\tilde{\beta})+bb']x_0',}$$

onde, $\boldsymbol{b = E(\beta-\tilde{\beta}) \approx 0}$.

Sendo assim, podemos escrever, usando algumas substituições, $\boldsymbol{\pi(x_i) = \pi_i}$ como:

$$\boldsymbol{\pi_i = P(Y_i=1) = \tilde{\pi}(x_i)+C_i,}$$

onde,

$\boldsymbol{C_i = (0,5-\tilde{\pi}(x_i))\tilde{\pi}(x_i)(1-\tilde{\pi}(x_i))x_i'V(\tilde{\beta})x_i}$ é o fator de correção. Este fator de correção é um estimador do vício da probabilidade de sucesso. (KING; ZENG, 2001)

A partir disso, temos que os estimadores KZ são definidos da seguinte maneira. O estimador KZ1 é o estimador da probabilidade de sucesso e é definido por:

$$\boldsymbol{\pi(x_i)^* = \tilde{\pi}(x_i)+C_i}$$

O estimador KZ2 é um estimador não viciado para a probabilidade de sucesso é dado por:

$$\boldsymbol{\pi(x_i)^{**} = \tilde{\pi}(x_i)-C_i}$$

2.3 Modelo Logito Limitado

As alternativas citadas anteriormente são maneiras de continuar utilizando o modelo logito usual, com algumas adaptações, para lidar com dados extremamente

desbalanceados. Existem outras alternativas que não seguem exatamente o modelo logito usual, mas o utilizam como base, como por exemplo, o modelo logito limitado.

O modelo Logito Limitado é uma modificação do modelo logito usual. Essa modificação consiste em um acréscimo de um parâmetro que limita superiormente a probabilidade do sucesso. O parâmetro de limitação ($\boldsymbol{\omega}$) tem a capacidade de absorver o impacto de possíveis covariáveis significativas excluídas da base de dados. (CRAMER, 2004)

A probabilidade de sucesso para este modelo, com o acréscimo do parâmetro $\boldsymbol{\omega}$ é dada por:

$$\boldsymbol{\pi_i = \omega \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}},} \tag{2.17}$$

Com $\boldsymbol{\omega}$ podendo variar entre 0 e 1 ($\mathbf{0 < \omega < 1}$).

2.3.1 Estimação

O logaritmo da Função de Verossimilhança é dado por:

$$\boldsymbol{l(\beta, \omega) = \sum_{i=1}^{n} \left(y_i \ln \left(\omega \left(\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \right) \right) \right) + (1 - y_i) \ln \left(1 - \omega \left(\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \right) \right) I_{(0,1)}(\omega)}$$

Uma vez que as probabilidades de sucesso e fracasso são dadas por: $\boldsymbol{P(Y_i = 1|x_i) = \pi_(x_i)}$ e $\boldsymbol{P(Y_i = 0|x_i) = 1 - \pi(x_i)}$, onde $\boldsymbol{Y_i \sim Bernoulli(\pi(x_i))}$ é a variável resposta.

Maximizando a expressão acima, obtemos os estimadores de máximo verossimilhança e as derivadas da função com relação aos parâmetros $\boldsymbol{\beta_0, \ldots, \beta_{1-p}}$ e $\boldsymbol{\omega}$ são:

$$\boldsymbol{\sum_{i=1}^{n} \omega \left(y_i - \pi(x_i) \right)}$$

$$\boldsymbol{\sum_{k=1}^{p-1} \sum_{i=1}^{n} x_{ij} \omega \left(y_i - \pi(x_i) \right)}$$

$$\boldsymbol{\sum_{i=1}^{n} \left[\frac{y_i - \pi(x_i)}{1 - \pi(x_i)} \right]}$$

Para a resolução destas equações é necessário utilizar um método de otimização para encontrar as estimativas de máxima verossimilhança, uma vez que essas equações são não-lineares nos parâmetros. Para melhorar lidar com a otimização, uma reparametrização é aconselhável, uma vez que do jeito que essas fórmulas estão, nem sempre é possível otimizá-las pelos métodos usuais, segundo Scacabarozi (2012). Com

a reparametrização $\boldsymbol{\theta = log\left(\frac{\omega}{1-\omega}\right)}$ a fórmula ficaria da seguinte maneira:

$$\boldsymbol{l\left(\beta,\omega\right) = \sum_{i=1}^{n}\left(y_i \ln\left(\left(\frac{e^{\theta}}{1+e^{\theta}}\right)\left(\frac{1}{1+e^{-x_i'\beta}}\right)\right)\right)+}$$
$$\boldsymbol{(1-y_i)\ln\left(1-\left(\frac{e^{\theta}}{1+e^{\theta}}\right)\left(\frac{1}{1+e^{-x_i'\beta}}\right)\right)}$$

onde $\boldsymbol{-\infty < \theta < \infty}$.

2.3.2 Interpretação dos Coeficientes

A interpretação dos coeficientes será feita considerando apenas uma covariável, categória, a fim de facilitar o entendimento. O caso para mais de uma covariável segue a mesma ideia.

Sendo x a covariável com dois níveis, ou seja, $\boldsymbol{x = 0}$ ou $\boldsymbol{x = 1}$, temos que:

$$\boldsymbol{\pi(1) = \omega\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \quad 1-\pi(1) = \frac{1+e^{\beta_0+\beta_1}(1-\omega)}{1+e^{\beta_0+\beta_1}}} \tag{2.18}$$

$$\boldsymbol{\pi(0) = \omega\frac{e_0^{\beta}}{1+e_0^{\beta}} \qquad 1-\pi(0) = \frac{1+e_0^{\beta}(1-\omega)}{1+e_0^{\beta}}} \tag{2.19}$$

Desta maneira, a razão de chances ($\boldsymbol{\Psi}$) é dada por:

$$\boldsymbol{\Psi = \frac{\left(\omega\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1+e^{\beta_0+\beta_1}}{1+(1-\omega)e^{\beta_0+\beta_1}}\right)}{\left(\omega\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1+e^{\beta_0}}{1+e^{\beta_0}(1-\omega)}\right)}}$$

Fazendo os devidos ajustes na fórmula acima temos que:

$$\boldsymbol{\Psi = \frac{e^{\beta_1}+(1-\omega)e^{\beta_0+\beta_1}}{1+(1-\omega)e^{\beta_0+\beta_1}}}$$

A interpretação deste resultado é: a chance do indivíduo ter o evento de interesse quando $\boldsymbol{x = 1}$ é $\frac{e^{\beta_1}+(1-\omega)e^{\beta_0+\beta_1}}{1+(1-\omega)e^{\beta_0+\beta_1}}$ a chance do indivíduo ter o evento de interesse quando $\boldsymbol{x = 0}$.

3 Simulação

Este capítulo tem como objetivo descrever o estudo de simulação realizado para reforçar a afirmação de King e Zeng (2001), onde se é falado que os estimadores de uma regressão logística com dados extremamente desbalanceados são viesados, ou seja, a presença de caudas pesadas atrapalha o funcionamento do modelo e que as probabilidades estimadas do modelo são subestimadas. No estudo proposto por King e Zeng (2001), são propostos duas alternativas para contornar a problemática apresentada, o método de correção a priori, o qual leva em consideração as informações a priori da população, e a "pesagem", onde o modelo e o cálculo do vício dos estimadores é corrigido por um peso, que desta maneira, reduz o viés dos estimadores e melhora as probabilidades estimadas.

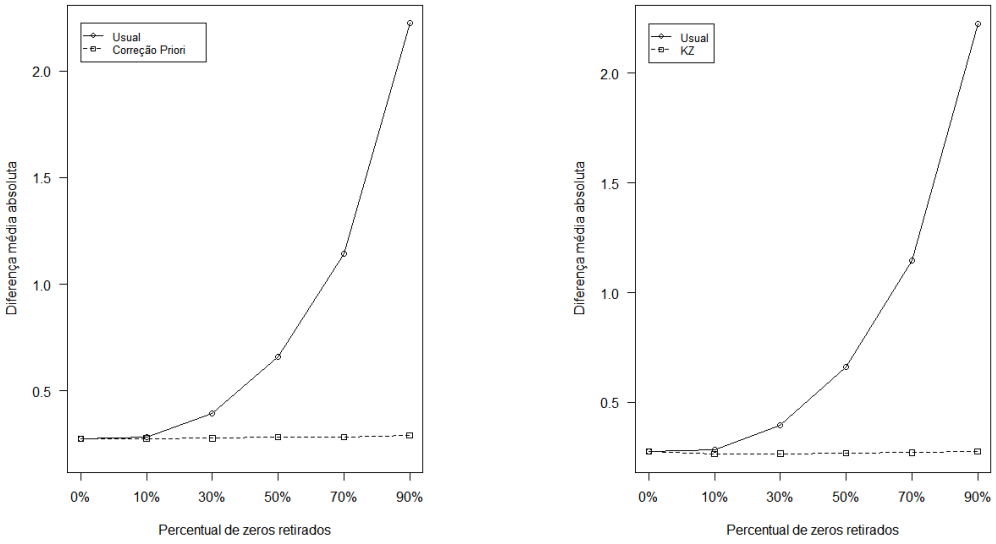
O breve estudo se baseia em comparar como ficam os estimadores antes e depois do ajuste proposto por King e Zeng (2001). Para a geração dos dados artificiais foi utilizado o modelo logístico usual, utilizando 1000 amostras de tamanho 1000, a fim de eliminar possíveis seleções de amostras que não refletem o comportamento geral do modelo, sendo a variável resposta com distribuição Bernoulli e apenas uma covariável com distribuição normal padrão. Os parâmetros pré-estabelecidos utilizados seguem a ideia do estudo proposto por Scacabarozi (2012), sendo eles $\boldsymbol{\beta_0 = -4.5}$ e $\boldsymbol{\beta_1 = 1}$.

Definidos os betas, o estudo de simulação se inicia com a geração dos 1000 modelos logísticos usuais e, como esperado, a estimação dos parâmetros não ficou tão ruim, sendo a média dos $\boldsymbol{\beta_0}$ igual a $\boldsymbol{-4.56}$ e dos $\boldsymbol{\beta_1}$ igual a $\boldsymbol{1.007}$, pois como os dados artificias utilizados no modelo foram gerados a partir dos betas iniciais falados anteriormente, é de se esperar que os gerados pelo modelo fiquem próximos dos reais. Sendo assim, este não é o foco do estudo de simulação. Se quer saber, com a amostragem *state-dependent*, como ficam as estimativas dos betas, ou seja, se de fato, ou então até que ponto, este tipo de amostragem é vantajoso para lidar com dados desbalanceados.

Como falado anteriormente, a amostragem do tipo *state-dependent* baseia-se na retirada de uma porcentagem de zeros, ou fracassos, da base de dados e em seguida, fazer a correção na matriz de pesos do modelo, levando em consideração a proporção de sucessos na amostra completa e na nova amostra, que teve retirada de zeros. Além da correção, a correção dos estimadores pelo vício traria benefícios, uma vez que a

variância dos estimadores sem o vício é menor do que a variância dos mesmos com vício McCullagh e Nelder (1989). Sendo assim, foram feitas 5 amostragens, retirando 10%, 30%, 50%, 70% e 90% dos zeros. Em cada uma destas amostragens, foram gerados os modelos usuais e foram gerados os modelos com a correção pelo peso, e nos com peso foram cálculos os vícios de cada estimador e retirados.

Para a melhor compreensão, foram gerados gráficos para mostrar o comportamento em cada reamostragem. Os gráficos têm como medida de comparação a diferença absoluta entre a média dos valores dos betas em cada retirada dos zeros, uma vez os estimadores podem ter sido maiores ou menores que o valor real de cada beta. Inicialmente, serão apresentados os gráficos para o $\boldsymbol{\beta_0}$, em seguida, serão apresentados os gráficos para o $\boldsymbol{\beta_1}$.



(a) Usual x Correção a priori

(b) Usual x KZ

Figura 1 – Gráficos para $\boldsymbol{\beta_0}$

Como podemos ver acima, para o $\boldsymbol{\beta_0}$, tanto os estimadores com o método de correção a priori, quanto os estimadores KZ apresentaram diferenças médias absolutas menores, em relação aos valores reais, do que os estimadores usuais, com os usuais podem a mais de 2 de diferença absoluta em relação ao valor real. Lembrando que o valor real para o $\boldsymbol{\beta_0}$ é igual a -4,5.

Percebendo o padrão de comportamento das duas correções, resolveu-se analisar também qual delas fica melhor, e por isso, o gráfico abaixo foi feito:

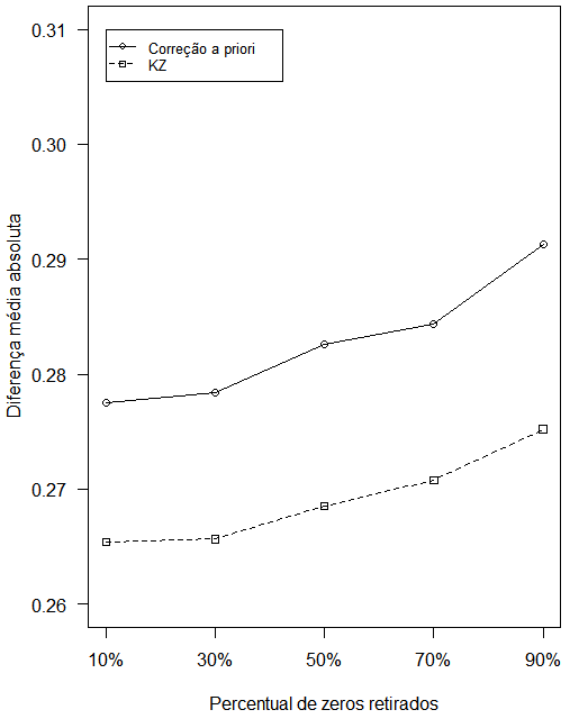


Figura 2 – $\boldsymbol{\beta_0}$ - Correção a priori x KZ

Como pode-se ver, a diferença entre os estimadores que sofreram correção pelo método de correção a priori estão mais distantes do que os estimadores que sofreram correção pelo peso e retirada do vício, mesmo que essa diferença não seja tão grande em média, os estimadores KZ apresentaram um comportamento melhor do que os com correção a priori para os $\boldsymbol{\beta_0}$.

Analisando agora o que acontece com o $\boldsymbol{\beta_1}$, também foi gerado um gráfico para facilitar o entendimento do comportamento deste estimador sem e com correção, Usual x KZ:

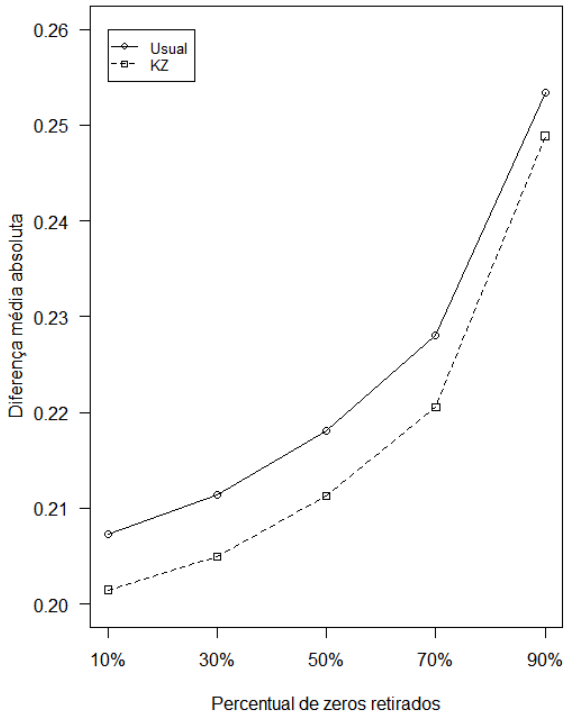


Figura 3 – $\boldsymbol{\beta_1}$ - Usual x KZ

Novamente, a diferença entre os estimadores gerados pelo modelo com a correção pelo peso e viés (KZ) ficam mais próximos dos valores reais do que os estimadores gerados pelo modelo sem essa correção.

Nota-se também, que conforme o percentual de zeros retirados da amostra aumenta, a diferença também aumenta, em ambos os casos. O mesmo acontece quando fala-se do $\boldsymbol{\beta_0}$, porém em proporções menores. Sendo assim, este estudo de simulação indica que a amostragem *state-dependent* é interessante, porém quando o percentual de zeros retirados da amostra é muito elevado, a correção não consegue minimizar os erros de maneira tão eficaz.

4 Resultados

Até agora foi muito falado sobre a regressão logística usual e as correções que deveriam ser feitas para torná-la uma boa opção quando os dados são extremamente balanceados. A partir de toda essa discussão feita até então, viu-se que a correção utilizando pesos e subtraindo o vício dos estimadores é a melhor opção para lidar com os dados da maneira proposta. Sendo assim, como os dados referentes aos acidentes de trânsito no Distrito Federal seguem o mesmo padrão de comportamento do que foi simulado e estudado anteriormente, optou-se por analisar estes dados de trânsito utilizando a regressão logística com as devidas correções apresentadas.

4.1 Variáveis

A base de dados fornecida pelo Detran-DF contém diversas variáveis que poderiam explicar a fatalidade de um acidente de trânsito no Distrito Federal, porém, foi feita uma pré-seleção de variáveis, as que de fato poderiam ter alguma relação com o evento, e depois foi testado no modelo de regressão logística. As variáveis pré-selecionadas foram: *Dia da semana, Horário do acidente, Tipo de envolvimento, Sexo do condutor, Idade do condutor, Tipo de veículo e se o condutor possuía ou não habilitação*. Abaixo, estas variáveis serão descritas com o intuito de melhor compreensão da base de dados e da programação utilizada para a geração dos resultados.

4.1.0.0.1 Dia da Semana

A variável Dia da Semana é uma variável categórica ordinal composta pelas seguintes categorias:

- 1 - Domingo
- 2 - Segunda-feira
- 3 - Terça-feira
- 4 - Quarta-feira
- 5 - Quinta-feira

- 6 - Sexta-feira
- 7 - Sábado

Nesta variável, foi feita a análise com uma alteração, dividindo os dias em dias úteis (segunda a sexta) e final de semana (sábado e domingo), porém, o ajuste do modelo se mostrou pior do que quando feito da maneira apresentada acima.

4.1.0.0.2 Horário do Acidente

A variável Horário do Acidente é uma variável categórica ordinal composta pelas seguintes categorias:

- Madrugada (Entre 00h e 05h59min)
- Manhã (Entre 06h e 11h59min)
- Tarde (Entre 12h e 17h59min)
- Noite (Entre 18h e 23h59min)

4.1.0.0.3 Tipo de Envolvimento

A variável Tipo de Envolvimento é uma variável categórica nominal que divide os condutores em três categorias, sendo elas: "Demais condutores", "Motociclista"e "Ciclista". A categoria "Demais condutores"é referente a todos os condutores de veículos de 4 rodas ou mais, ou seja, automóveis, ônibus, caminhões, etc.

4.1.0.0.4 Sexo do Condutor

A variável Sexo do Condutor é uma variável categórica nominal, podendo o condutor ser do sexo masculino ou do sexo feminino.

4.1.0.0.5 Idade do Condutor

A variável Idade do Condutor é uma variável quantitativa contínua e representa a idade do condutor no momento do acidente.

4.1.0.0.6 Tipo de Veículo

A variável Tipo de Veículo é uma variável categórica nominal e representa o veículo causador do acidente. Os tipos de veículo são: Automóvel, Bicicleta, Caminhão, Caminhonete, Carroça, Charrete*, Microônibus, Moto, Ônibus, Outro**

** A categoria Charrete foi removida dos cálculos por apresentar apenas uma observação e possuir dados faltantes em várias variáveis.*

*** Outro: Pode ser qualquer veículo que não citado anteriormente.*

4.1.0.0.7 Não Habilitado

A variável Não Habilitado é uma variável categórica nominal e foi utilizada, no banco de dados e nos cálculos, como uma variável dicotômica, podendo ser igual a 0 ou 1, sendo:

- 0 - Motorista Habilitado
- 1 - Motorista Não-Habilitado

4.2 Distrito Federal

Apesar de terem sido apresentadas alternativas para contornar o problema da baixa proporção de sucessos, as correções propostas por King e Zeng (2001) não foram necessárias neste modelo inicial, pois por possuir um grande número de observações, quase 40 mil, percebeu-se que a baixa proporção de sucessos não impactou tão negativamente quanto esperado na aplicação do modelo logístico, na verdade, mostrou-se um melhor ajuste na regressão logística usual do que na regressão logística com correção, utilizando o teste de Hosmer e Lemeshow. O modelo logito limitado, proposto por Cramer (2004) teve um bom comportamento com os dados, se mostrando uma alternativa viável para lidar com este tipo de situação. Os valores dos coeficientes e o teste de Hosmer e Lemeshow mostrou que os dois modelos tiveram resultados bem próximos. Mais a frente serão comparados os dois modelos.

Sendo assim, pela maior facilidade em utilizar o modelo usual em relação ao limitado, optou-se pela utilização do usual para representar os acidentes em todo o Distrito Federal.

A base de dados para os acidentes do Distrito Federal é composta por 31329 observações, sendo apenas 1036 destes fatais.

4.2.1 Seleção das Variáveis

Inicialmente, foi construído um modelo para todos os acidentes no Distrito Federal com todas as variáveis listadas acima para poder utilizar um critério de seleção a partir delas. O método utilizado foi o *Stepwise* uma vez que esse é um dos métodos de seleção de variáveis mais utilizados em regressão logística. O método é uma mistura entre inclusão e exclusão de variáveis explicativas. (PAULA, 2013)

Sendo assim, o modelo foi rodado com todas as variáveis listadas e a partir do método de seleção *Stepwise* e como falado anteriormente, na seção 3.1.5, o objetivo é obter o modelo com o menor AIC possível, sendo assim, segue os valores do AIC dos modelos e suas respectivas desviâncias:

Tabela 1 – AIC - Regressão Logística Usual

Variáveis no modelo	**G.L**	***Deviance***	**AIC**
(-) Tipo de Envolvimento	2	7894	7938
(-) Idade	1	7896	7938
(-) Dia da Semana	6	7929	7961
(-) Não Habilitado	1	7920	7962
(-) Sexo	1	7945	7987
(-) Horário	3	7989	8027
(-) Tipo de Veículo	9	8031	8057

Pode-se observar na tabela que os modelos tirando a variável Tipo de Envolvimento e a variável Idade possuem o mesmo valor no AIC, porém, como critério de desempate, foi escolhido a desviância para fazer a seleção do modelo mais apropriado e, como o modelo sem a variável Tipo de Envolvimento possui desviância menor do que o modelo sem a variável Idade, optou-se pelo primeiro modelo.

4.2.2 Modelo

Sendo assim, o modelo final obtido é composto pelas variáveis Idade, Dia da Semana, Sexo, Horário, Tipo de Veículo e Não Habilitado, ou seja, tirando apenas a variável Tipo de Envolvimento. Os resultados obtidos a partir do modelo usual final são:

Tabela 2 – Resumo - Regressão Logística Usual

Coeficiente	**Valor**	**Erro Padrão**	**Z-valor**	**P-valor**
Intercepto	-3,184	0,176	-18,060	<0.001*
Idade	0,004	0,003	1,429	0.1529
Dia:Segunda	-0,393	0,126	-3,117	0.0018*
Dia:Terça	-0,492	0,129	-3,804	<0.001*
Dia:Quarta	-0,586	0,134	-4,380	<0.001*
Dia:Quinta	-0,484	0,129	-3,756	<0.001*
Dia:Sexta	-0,220	0,115	-1,918	0.0551
Dia:Sábado	-0,107	0,111	-0,968	0.3329
Não Habilitado:1	0,849	0,151	5,624	<0.001*
Sexo:Masculino	0,770	0,117	6,580	<0.001*
Horário:Manhã	-1,001	0,118	-8,483	<0.001*
Horário:Noite	-0,544	0,112	-4,848	<0.001*
Horário:Tarde	-1,030	0,116	-8,917	<0.001*
Veículo:Bicicleta	0,487	0,150	3,257	0.0011*
Veículo:Caminhão	1,420	0,138	10,300	<0.001*
Veículo:Caminhonete	0,297	0,149	2,001	0.0454*
Veículo:Carroça	1,651	0,769	2,148	0.0318*
Veículo:Microônibus	0,323	0,461	0,701	0.4835
Veículo:Moto	-0,222	0,085	-2,628	0.0086*
Veículo:Ônibus	0,698	0,149	4,677	<0.001*
Veículo:Outro	2,328	0,840	2,771	0.0056*

De acordo com o modelo acima, diversas variáveis se mostraram significativas se levar em consideração um nível de significância $\boldsymbol{\alpha}$ igual a 5%, com algumas exceções, como por exemplo, a idade. Ou seja, uma rápida conclusão que pode-se tirar deste modelo é que não há evidências estatísticas suficientes para dizer que a idade influencia em um acidente ser fatal ou não no Distrito Federal entre os anos de 2015 e 2016.

4.2.3 Interpretação dos Coeficientes

Esta seção é destinada para a interpretação dos resultados obtidos pelo modelo proposto para os dados do Distrito Federal. Vale ressaltar que a razão de chances obtida é dada tal que as demais variáveis sejam constantes.

Dia da Semana: Nesta variável, dividida em 7 categorias, tendo Domingo como referência, podemos perceber que os dias da semana, com exceção de Sexta, se mostraram significativos quando comparados com Domingo, e todos com uma chance

menor de ocorrer acidente fatal do que Domingo. Por exemplo, se selecionarmos a Quarta, vemos que a chance de se ter acidente fatal na Quarta é ($\boldsymbol{e^{\beta} = e^{-0,586}}$) 0,5565 vezes a chance de se ter um acidente fatal no Domingo, ou seja, a chance de se ter acidente fatal na quarta-feira é 44,35% menor do que no domingo. A interpretação para as demais categorias ocorre de maneira similar.

Horário: Nesta variável ocorre o mesmo que ocorreu com a variável anterior. Nota-se que todas as categorias foram significativas quando comparadas com a de referência que é de Madrugada. Como os valores foram menores do que zero, implica dizer que a chance de se ter um acidente fatal de madrugada é superior as chances de se ter de manhã, tarde ou noite. Por exemplo, a chance de se ter acidente fatal de manhã é ($\boldsymbol{e^{\beta} = e^{-1,001}}$) 0,3675 vezes a chance de se ter um acidente fatal de madrugada, ou seja a chance de se ter acidente fatal pela manhã é 63,25% menor do que de madrugada. A interpretação para as demais categorias ocorre de maneira similar.

Sexo: Nesta variável, dividida em duas categorias, Masculino e Feminino, ocorre o mesmo que nas variáveis anteriores, onde há evidências estatísticas suficientes para afirmar que a chance de se ter um acidente fatal quando o condutor é do sexo masculino é maior do que se o condutor for do sexo feminino. Mais especificamente, a chance de se ter um acidente fatal quando o condutor é do sexo masculino é ($\boldsymbol{e^{\beta} = e^{0,77}}$) 2,16 vezes a chance de se ter um acidente fatal quando o condutor é do sexo feminino.

Idade: Na variável idade, não há evidências estatísticas suficientes para afirmar que a idade influencie ou não no acidente ser fatal ou não.

Veículo: Esta variável segue os mesmos princípios das variáveis Sexo, Horário e Dia da semana, porém, nem todas as categorias foram significativas estatisticamente. Apenas a categoria Microônibus não foi significativa. Nas que foram significativas, com exceção da Moto a chance de se ter um acidente fatal é maior nessas categorias do que na categoria de referência que é Automóvel.

Não Habilitado: Esta variável segue os mesmos princípios das variáveis Sexo, Horário, Dia da semana e Veículo, onde há evidências estatísticas suficientes para afirmar que a chance de se ter um acidente fatal quando o condutor não possui habilitação é maior do que a chance de se ter um acidente fatal quando o condutor é habilitado. Quando o condutor não é habilitado, essa chance é 2,3 vezes a chance de quando o condutor é habilitado, ou seja, 230% maior.

4.2.4 Diagnóstico

Achado o melhor modelo com base no AIC e Deviance, testará-se agora para ver se fato o modelo proposto consegue explicar bem o que se observa. Para isso, foi escolhido o teste de Hosmer e Lemeshow. Sendo assim, segue os resultados do teste e a relação de observados e esperados pelo modelo a fim de comparação:

Tabela 3 – Teste HosmerLemeshow - Informações

Estatística Qui-Quadrado	Graus de Liberdade	P-valor
7,1	8	0,5

Tabela 4 – Teste HosmerLemeshow - Números Esperados e Observados de acidentes fatais no DF

Maior probabilidade estimada da classe	**Esperado**	**Observado**
.0136	31	34
.0172	45	50
.0202	53	53
.0232	62	50
.0265	70	70
.0306	81	68
.0355	94	91
.0442	112	120
.0589	143	154
.0497	255	255

Como pode-se notar na Tabela 3, não há evidências estatísticas para rejeitar a hipótese nula, sendo assim, diz-se que o modelo de fato explica bem o que se observa. Já olhando para a Tabela 4, vê-se que o modelo conseguiu explicar exatamente o que é observado em várias das divisões das probabilidades estimadas. Em algumas houve subestimação e em outras uma sobrestimação, porém, isso não impactou no p-valor obtido pelo teste.

Para complementar o diagnóstico, foi gerado um quadro composto por quatro gráficos para auxiliar a visualização da qualidade do modelo. Estes gráficos encontram-se abaixo:

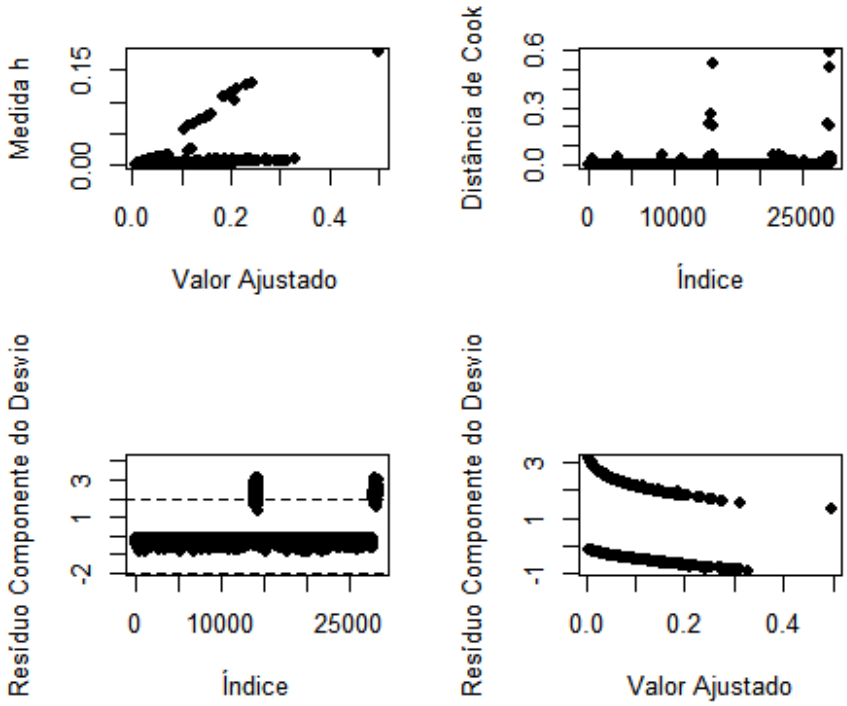


Figura 4 – Gráficos de Diagnóstico

Pelo gráfico da distância de Cook, vemos que têm-se apenas alguns pontos influentes, o que é um bom indício para o modelo. A distância de Cook é utilizada para medir a influência das observações nas estimativas dos coeficientes. (PAULA, 2013)

Optou-se pela utilização de gráficos para interpretar o que acontece com os resíduos pela dificuldade em saber quais pontos estão mais afastados no subespaço gerado pelos resíduos quando a probabilidade ajustada é muito alta ou baixa. A escolha deste resíduo, ao invés do resíduo usual, se dá pela sua distribuição, que segundo estudos de simulação não se afasta muito da distribuição normal padrão, Paula (2013). O gráfico do resíduo componente do desvio pode informar, assim como o da medida h, emparelhamentos discrepantes com algum tipo de influência nos resultados do modelo. (PAULA, 2013)

Pelos gráficos, têm-se que apenas um ponto se mostrou influente, porém não há indícios de observações que destoam muito das suposições da distribuição Bernoulli para a resposta, os demais se mostraram dentro dos conformes.

4.2.5 Comparação Logito Limitado

Esta seção destina-se para comparar os resultados obtidos pelo modelo logístico limitado proposto por Cramer (2004) com o modelo logístico usual para os dados do Distrito Federal.

Como primeira forma de comparação, serão apresentados os valores dos coeficientes de cada modelo na tabela abaixo:

Tabela 5 – Resumo - Modelo Usual X Modelo Limitado

Coeficiente	**Usual**	**Limitado**
Intercepto	-3,1842	-2,6848
Idade	0,0039	0,0040
Dia:Segunda	-0,3930	-0,4119
Dia:Terça	-0,4919	-0,5104
Dia:Quarta	-0,5859	-0,6089
Dia:Quinta	-0,4839	-0,5006
Dia:Sexta	-0,2204	-0,2359
Dia:Sábado	-0,1075	-0,1131
Não Habilitado:1	0,8485	0,8853
Sexo:Masculino	0,7701	0,7824
Horário:Manhã	-1,0013	-1,0382
Horário:Noite	-0,5442	-0,5689
Horário:Tarde	-1,0302	-1,0669
Veículo:Bicicleta	0,4871	0,5017
Veículo:Caminhão	1,4202	1,4930
Veículo:Caminhonete	0,2971	0,3059
Veículo:Carroça	1,6507	1,7536
Veículo:Microônibus	0,3232	0,3376
Veículo:Moto	-0,2223	-0,2280
Veículo:Ônibus	0,6980	0,7218
Veículo:Outro	2,3277	2,4288

O valor obtido para o parâmetro de limitação da razão de chances, $\boldsymbol{\omega}$, foi 0,6377.

Como pode-se observar, os valores dos coeficientes no modelo usual e no modelo limitado ficaram bem próximos, com o modelo limitado tendo valores um pouco superiores em relação ao usual.

Em seguida, será comparado os resultados obtidos pelo teste de Hosmer e Lemeshow para os dois modelos, junto com a tabela de observados x esperados de cada modelo.

Tabela 6 – Teste HosmerLemeshow - Informações

Modelo	Estatística Qui-Quadrado	Graus de Liberdade	P-valor
Usual	7,1	8	0,5
Limitado	7,1	8	0,5

Tabela 7 – Teste HosmerLemeshow - Números Esperados e Observados de acidentes fatais no DF para os modelos usual e limitado

Classes	Modelo Usual		Modelo Limitado	
	Esperado	Observado	Esperado	Observado
1ª Classe	34	31	34	30
2ª Classe	50	45	50	45
3ª Classe	53	53	52	52
4ª Classe	50	62	51	62
5ª Classe	70	70	70	70
6ª Classe	68	81	66	81
7ª Classe	91	94	93	94
8ª Classe	120	112	121	112
9ª Classe	154	143	153	144
10ª Classe	255	255	255	255

Como pode-se observar pelo teste de Hosmer e Lemeshow, os dois modelos ficaram muito próximos. Obtiveram a mesma estatística qui-quadrado no teste e os valores esperados para acidentes fatais no DF ficou muito próximo. Sendo assim, não se vê muita vantagem em utilizar o modelo limitado, pois em relação ao usual, obteve resultados resultados muito semelhantes e o esforço computacional para gerar este modelo é consideravelmente superior ao modelo usual.

4.3 BRs

Nesta seção, será mostrará-se os resultados obtidos para uma menor quantidade de observações. Para isso, foram escolhidas apenas as observações referentes às BRs que cruzam o Distrito Federal. Os resultados obtidos serão apresentados sob as três óticas abordadas neste trabalho: Modelo Usual, Modelo Usual com as correções propostas por King e Zeng (2001) e Modelo Limitado proposto por Cramer (2004), porém, **novamente foi escolhido o modelo usual para representar os resultados obtidos e fazer o diagnóstico**, uma vez que novamente o modelo

limitado ficou muito semelhante ao usual e as correções não trouxeram benefícios para o estudo, entretanto, será evidenciado os resultados dos três modelos.

A base de dados dos acidentes para as BRs do Distrito Federal é composta por 1835 observações, sendo apenas 221 delas são de acidentes fatais.

4.3.1 Seleção de Variáveis

Assim como feito para todos os acidentes do Distrito Federal, foi construído um modelo para todos os acidentes registrados nas BRs com as mesmas variáveis listadas anteriormente e adicionando uma nova variável, Habilitação Suspensa, que representa se o motorista estava ou não com a habilitação suspensa no momento do acidente, para poder utilizar um critério de seleção a partir delas. O método utilizado foi novamente o *Stepwise*.

Sendo assim, o modelo foi rodado com todas as variáveis listadas e a partir do método de seleção *Stepwise* e o procedimento foi semelhante ao utilizado para os dados do DF. Com isso, o modelo final ficou composto pelas variáveis Horário, Tipo de Veículo e Habilitação Suspensa, com AIC igual a 398 e *Deviance* igual a 386. Os demais modelos, com mais ou menos variáveis, tiveram valores AIC e *Deviance* superiores aos citados anteriormente.

4.3.2 Modelo

Como falado anteriormente, o modelo final ficou composto pelas variáveis Horário, Tipo de Veículo e Habilitação Suspensa. Os resultados obtidos a partir do modelo usual final são:

Tabela 8 – Resumo - Regressão Logística Usual

Coeficiente	**Valor**	**Erro Padrão**	**Z-valor**	**P-valor**
Intercepto	-1,456	0,230	-6,334	<0.001*
Horario::Manhã	-1,069	0,289	-3,693	<0.001 *
Horario::Noite	-0,271	0,264	-1,027	0,3046
Horario::Tarde	-1,084	0,281	-3,855	<0.001*
Veículo::Caminhão	1,790	0,288	6,224	<0.001*
Veículo::Caminhonete	0,762	0,292	2,606	0.0092*
Veículo::Moto	-0,528	0,225	-2,348	0.0189*
Veículo::Ônibus	1,340	0,424	3,159	0.0016*
Habilitação Suspensa::1	2,578	0,935	2,756	0.0058*

4.3.3 Interpretação dos Coeficientes

Esta seção é destinada para a interpretação dos resultados obtidos pelo modelo proposto para os dados das BRs do Distrito Federal. Vale ressaltar que a razão de chances obtida é dada tal que as demais variáveis sejam constantes.

A interpretação dos coeficentes do modelo apresentado anteriormente se dá da mesma maneira da interpretação feita para o modelo do Distrito Federal. o exponencial do $\boldsymbol{\beta_i}$ representa a razão de chances da categoria tida como referência e a classe estudada. Dessa maneira, pode-se concluir algumas coisas, como por exemplo, a chance de se ter um acidente fatal de madrugada nas BRs segue o mesmo padrão do DF, onde é maior de madrugada do que de manhã e de tarde, porém não se pode afirmar que o mesmo acontece para a noite, uma vez que não há evidências estatísticas para afirmar isso. A interpretação é a mesma para o tipo de veículo e habilitação suspensa.

4.3.4 Diagnóstico

Achado o melhor modelo com base no AIC e Deviance, testará-se agora para ver se fato o modelo proposto consegue explicar bem o que se observa. Para isso, foi escolhido o teste de Hosmer e Lemeshow. Sendo assim, segue os resultados do teste e a relação de observados e esperados pelo modelo a fim de comparação:

Tabela 9 – Teste HosmerLemeshow - Informações

Estatística Qui-Quadrado	Graus de Liberdade	P-valor
5,6	8	0,7

Tabela 10 – Teste HosmerLemeshow - Números Esperados e Observados de acidentes fatais nas BRs

Maior probabilidade estimada da classe	**Esperado**	**Observado**
.0451	16	13
.0731	22	18
.0741	9	14
.0949	7	12
.151	51	48
.189	16	16
.701	46	46

Analisando as informações acima, nota-se que não houve rejeição da hipótese nula no teste de Hosmer e Lemeshow, indicando assim, que esse modelo pode estar de fato ajustado bem para os dados. Porém, apenas essa informação não é suficiente. É necessário sabermos mais sobre os resíduos, sendo assim, abaixo estão os gráficos que indicam sobre pontos influentes e o comportamento dos resíduos, para que seja analisado mais a fundo o modelo:

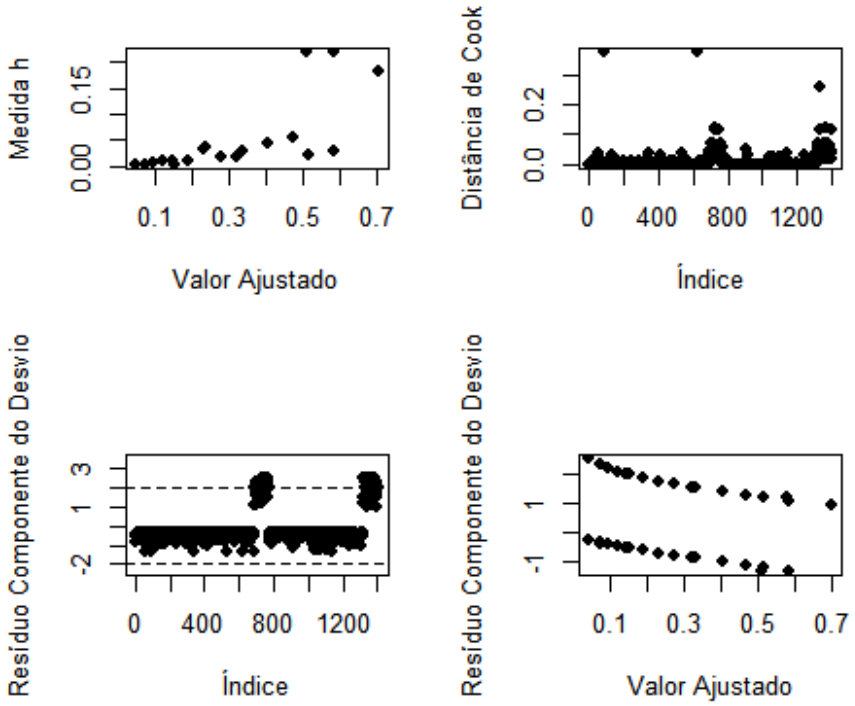


Figura 5 – Gráficos de Diagnóstico

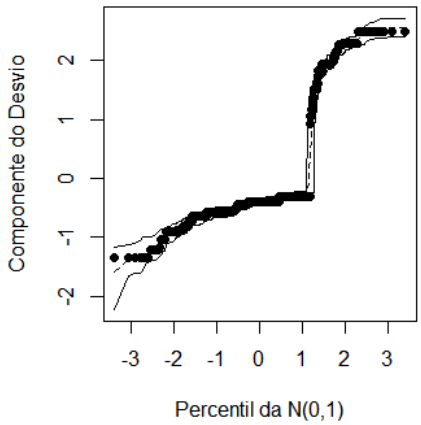


Figura 6 – Gráficos de Diagnóstico

Pelo gráfico da distância de Cook, vemos que têm-se apenas alguns pontos influentes, o que é um bom indício para o modelo. A distância de Cook é utilizada para medir a influência das observações nas estimativas dos coeficientes. (PAULA, 2013)

Optou-se pela utilização de gráficos para interpretar o que acontece com os resíduos pela dificuldade em saber quais pontos estão mais afastados no subespaço gerado pelos resíduos quando a probabilidade ajustada é muito alta ou baixa. A escolha deste resíduo, ao invés do resíduo usual, se dá pela sua distribuição, que segundo estudos de simulação não se afasta muito da distribuição normal padrão. O gráfico do resíduo componente do desvio pode informar, assim como o da medida h, emparelhamentos discrepantes com algum tipo de influência nos resultados do modelo. (PAULA, 2013)

Pelos gráficos, têm-se que poucos pontos se mostraram influentes, porém não há indícios de observações que destoem muito das suposições da distribuição Bernoulli para a resposta, os demais se mostraram dentro dos conformes.

O gráfico de envelope, representado na Figura 6, indica como os resíduos deveriam se comportar, considerando a média e os intervalos de confiança, superiores e inferiores. Sendo assim, como todos os pontos ficaram dentro dos limites, é um ótimo indício para mostrar que o modelo está de fato se ajustando bem aos dados.

4.3.5 Comparação Logito Limitado e KZ

Esta seção destina-se para comparar os resultados obtidos pelo modelo logístico limitado proposto por Cramer (2004), pelo modelo usual com as correções propostas por King e Zeng (2001) e o modelo logístico usual para os dados das BRs do Distrito Federal.

Como primeira forma de comparação, serão apresentados os valores dos coeficientes de cada modelo na tabela abaixo:

Tabela 11 – Resumo - Modelo Usual X Modelo Limitado

Coeficiente	**Usual**	**Limitado**	**KZ**
Intercepto	-1,456	-1,456	-1,291
Horário::Manhã	-1,069	-1,069	-1,264
Horário::Noite	-0,271	-0,271	-0,465
Horário::Tarde	-1,084	-1,084	-1,287
Veículo::Caminhão	1,790	1,790	1,871
Veículo::Caminhonete	0,762	0,762	1,123
Veículo::Moto	-0,528	-0,528	-0,521
Veículo::Ônibus	1,340	1,340	1,613
Habilitação Suspensa::1	2,578	2,578	1,942

O valor obtido para o parâmetro de limitação da razão de chances, ω, no modelo logito limitado foi de aproximadamente 0,9998.

Como pode-se notar, os valores do modelo usual e limitado ficaram iguais com o arredondamento, já os valores para o modelo com a correção KZ ficou consideravelmente diferente dos demais.

Só com essas informações não é o bastante para se avaliar qual modelo ficou de fato melhor, portanto, foi feito o teste de Hosmer e Lemeshow para todos os modelos, que encontra-se a seguir:

Tabela 12 – Teste HosmerLemeshow - Informações

Modelo	Estatística Qui-Quadrado	Graus de Liberdade	P-valor
Usual	7,1	8	0,5
Limitado	7,1	8	0,5
KZ	59	8	<0,001

Como pode-se observar, a estatística qui-quadrado do modelo usual e do limitado ficaram iguais. Com a correção KZ a estatística do teste apresentou um valor superior e a hipótese nula foi rejeitada. Desta forma, pode-se afirmar que o o modelo com as correções propostas por King e Zeng (2001) não trouxeram melhorias para o modelo. Justamente por isso, optou-se pela utilização do modelo usual para explicar o fenômeno estudado, uma vez que para o desenvolvimento do limitado é necessário previamente rodar o usual e depois aplicar as mudanças para se tornar o logito limitado, se tornando assim mais exaustivo computacionalmente, e o modelo com as correções ficou consideravelmente pior que os outros, sendo assim, descartado como uma opção para estes dados.

5 Conclusão

Dado o exposto nas seções anteriores, viu-se que a correção pelo método de King e Zeng (2001) não trouxe vantagens em relação ao modelo usual, principalmente pela grande quantidade de observações presentes no banco de dados. Já o modelo logito limitado, proposto por Cramer (2004), se mostrou semelhante ao usual, obtendo até a mesma estatística qui-quadrado no teste de Hosmer e Lemeshow para os dados de todo o Distrito Federal, porém, pelo seu alto esforço computacional exigido, talvez não seja a melhor opção para lidar com essa situação, uma vez que o usual se encaixou bem com o fenômeno estudado e, por isso, optou-se por usar o modelo usual para explicar os principais fatores que influenciam na fatalidade dos acidentes fatais no trânsito do Distrito Federal em 2015 e 2016.

Em vista dos resultados obtidos, algumas informações interessantes sobre o trânsito do Distrito Federal. A mais interessante delas, é sobre a fatalidade envolvendo motos e carros. É de se esperar, pelo menos o senso comum aborda isso, que é mais "fácil"morrer dirigindo uma moto do que um carro, porém, os dados mostraram, que para o DF entre 2015 e 2016 isso não é verdade, na verdade, a chance de se ter uma fatalidade quando uma moto está envolvida é menor do que quando um carro está envolvido. Outros resultados ficaram na medida do esperado, como a chance do acidente ser fatal de madrugada é superior à chance do acidente ser fatal de manhã, de tarde e de noite. O mesmo para os dias da semana, onde a chance é maior no domingo do que nos demais dias da semana, porém o mesmo não se pode falar entre domingo e sábado e domingo e sexta.

Os resultados obtidos para as BRs do Distrito Federal mostraram que os modelos usual e limitado explicam bem o que acontece, entretanto, pela maior complexidade em utilização do modelo limitado, o modelo usual foi o escolhido para ser o representante deste corte da base de dados estudada. As interpretações são as mesmas utilizadas no modelo para o Distrito Federal, porém com as mudanças da variáveis citadas anteriormente, as quais no modelo final só estiveram três presentes, o horário, o tipo de veículo e se o motorista estava ou não com a habilitação suspensa.

Referências

ALVEZ, P. F.; SILVA, A. R. da. Modelagem de eventos raros: Uma aplicação utilizando regressão probit. *Revista da Estatística UFOP*, v. 6, n. 6, 2017. Nenhuma citação no texto.

CANNEL, A.; WRIGHT, C. L. *Reduzindo Acidentes - O papel da fiscalização de trânsito e do treinamento de motoristas*. [S.l.]: Banco Interamericano de Desenvolvimento, 2000. v. 1. Nenhuma citação no texto.

CRAMER, J. Scoring bank loans that may go wrong - a case study. *Tinbergen Institute Amsterdan and Rotterdam*, v. 1, n. 1, 2004. Citado 8 vezes nas páginas 18, 19, 33, 41, 47, 48, 53 e 55.

DETRAN. *DF entra no oitavo mês de redução de mortes no trânsito*. 2017. Website. Disponível em: <http://detran.df.gov.br/noticias/item/3244-df-entra-no-oitavo-m%C3%AAs-de-redu%C3%A7%C3%A3o-de-mortes-no-tr%C3%A2nsito.html>. Acesso em: 01 set 2017. Citado na página 17.

EVANS, L. *Traffic Safety and the Driver*. [S.l.]: Van Nostrand Reinhold, 1991. v. 1. Nenhuma citação no texto.

FORBES, T. *Human Factors in Highway Traffic Safety Research*. [S.l.]: Wiley-Interscience, John Wiley & Sons, 1972. v. 1. Nenhuma citação no texto.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. 2. ed. [S.l.]: John Wiley & Sons, 2000. Citado 5 vezes nas páginas 20, 21, 26, 28 e 29.

IPEA, E. Estimativa dos custos dos acidentes de trânsito no brasil. *IPEA*, v. 1, n. 1, 2016. Citado na página 17.

KING, G.; ZENG, L. Logistic regression in rare events data. *Political Analysis*, v. 9, n. 2, p. 137–163, 2001. Citado 11 vezes nas páginas 18, 19, 29, 31, 32, 35, 41, 48, 53, 54 e 55.

KUME, L.; NERI, M. É possível reduzir as mortes no trânsito? *FGV*, v. 1, n. 1, 2007. Nenhuma citação no texto.

LIMA, D. D.; RODRIGUES, J. *Educação para o trânsito no Ensino Médio*. [S.l.]: David Duarte Lima, 2016. v. 1. Citado na página 17.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. [S.l.]: CRC press, 1989. v. 37. Citado 4 vezes nas páginas 19, 22, 31 e 36.

PAULA, G. A. Modelos de regressão com apoio computacional. *Universidade de São Paulo*, v. 1, n. 1, 2013. Citado 6 vezes nas páginas 27, 28, 42, 46, 52 e 53.

RODRIGUES, A. S. *Regressão Logística com erro de medida: comparação de métodos de estimação*. Dissertação (Dissertação de Mestrado) — Universidade de São Paulo, 2013. Nenhuma citação no texto.

SCACABAROZI, F. N. *Modelagem de eventos raros:um estudo comparativo.* Dissertação (Dissertação de Mestrado) — Universidade Federal de São Carlos, 2012. Citado 3 vezes nas páginas 18, 33 e 35.

STUKEL, T. A. Generalized logistic models. *Journal of the American Statistical Association*, Taylor & Francis, v. 83, n. 402, p. 426–431, 1988. Citado 3 vezes nas páginas 18, 19 e 29.

TELES, A. B. C. *Trânsito do Distrito Federal*. [S.l.]: Antônio Bomfim Carvalho Teles, 2002. v. 1. Nenhuma citação no texto.